Received: 1 June 2025 | Revised: 25 September 2025 | Accepted: 26 September 2025

DOI: 10.1002/alz.70886

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

RESEARCH ARTICLE

Integrative multi-omics approaches identify molecular pathways and improve Alzheimer's disease risk prediction

Rasika Venkatesh[1,2] | **Katie M. Cardone**[2] | **Yuki Bradford**[2] | **Anni K. Moore**[1,2] | **Rachit Kumar**[1,2,3] | **Jason H. Moore**[4] | **Li Shen**[2,5] | **Dokyoon Kim**[2,6] | **Marylyn D. Ritchie**[2,5,6]

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[4]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, California, USA

[5]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[6]Division of Informatics, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Rasika Venkatesh, Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
Email: rasika.venkatesh@pennmedicine.upenn.edu

Marylyn D. Ritchie, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
Email: marylyn@pennmedicine.upenn.edu

Funding information
the National Institutes of Health, Grant/Award Number: R01HL169458,U01AG066833,P30AG073105; National Human Genome Research Institute, Grant/Award Number: T32HG000046; National Institute on Aging, Grant/Award Numbers: U19 AG074879, P30AG073105; NHLBI Division of extramural Research, Grant/Award Number: HL-169458

Abstract

INTRODUCTION: Alzheimer's disease (AD) is a complex neurodegenerative disorder with heterogeneous genetic and molecular underpinnings. Polygenic scores (PGS) capture little of this complexity.

METHODS: We conducted genome-, transcriptome-, and proteome-wide association studies (G/T/PWAS) on 15,480 individuals from the Alzheimer's Disease Sequencing Project R4 (ADSP) to identify AD-associated signals, followed by pathway enrichment analysis. Integrative risk models (IRMs) were developed using genetically regulated components of gene and protein expression and clinical covariates. Elastic-net logistic regression and random forest classifiers were evaluated using standard metrics and compared against baseline PGS.

RESULTS: Known and novel signals were identified via G/T/PWAS. Enrichment analyses highlighted cholesterol and immune signaling pathways. The best-performing IRM, random forest with transcriptomic and covariate features, achieved area under the receiver operating characteristic (AUROC) of 0.703 and area under the precision-recall curve (AUPRC) of 0.622, significantly outperforming PGS and baseline models.

DISCUSSION: Integrating univariate discovery approaches with multivariate modeling enhances AD risk prediction and offers novel insights into underlying biological processes.

KEYWORDS
Alzheimer's disease, association study, machine learning, multi-omics integration, pathway enrichment, polygenic risk scores, risk prediction, statistical genetics

Highlights

- Identified novel contributions to Alzheimer's disease (AD) from a multi-omics perspective.
- Integrated genome-wide association studies (GWAS), transcriptome-wide association studies (TWAS), and proteome-wide association studies (PWAS) in a unified association study framework.
- Developed a method for predicting heritable risk of late-onset AD.
- Demonstrated that ancestry-aware modeling improves AD risk prediction accuracy.

1 | BACKGROUND

Alzheimer's disease (AD) is the most prevalent neurodegenerative condition affecting the aging population, posing a significant public health challenge as life expectancy continues to rise.[1,2] Late-onset AD (LOAD), the most common form of AD, has proven particularly challenging to characterize in terms of genetic risk.[1] Much of the genetic contribution to LOAD remains unexplained, complicating efforts to develop predictive models.[2] Most existing models rely on genetic variants identified through genome-wide association studies (GWAS). The apolipoprotein E (*APOE*) ε4 allele is the most well-established genetic risk factor for AD, accounting for one-quarter of the heritable contributions to liability, with total heritability of AD estimated to be between 58% and 75%.[3–5] The contributions of variants in other genes, such as *BIN1*, have also been characterized via large-scale GWAS.[4,6]

Emerging evidence indicates that AD-associated genetic variants converge on key biological pathways. These include cholesterol and lipid metabolism, neuroinflammation, and synaptic function.[7] Disruptions in these processes contribute to hallmark AD pathologies such as amyloid-β accumulation, tau hyperphosphorylation, and neuronal loss.[7] This biological context has spurred a shift toward approaches that use transcriptomic, proteomic, and clinical data with genomic information to uncover the molecular mechanisms underlying AD. Recent evidence from other complex traits suggest that tissue-specific gene expression and protein expression derived from transcriptome- and proteome-wide association studies (TWAS and PWAS, respectively) provide orthogonal and complementary information to genetic data alone, offering opportunities to improve risk prediction models and identify putatively novel genes and proteins.[8,9] TWAS have identified 54 hippocampal genes linked to AD risk, with fine-mapping prioritizing 24 candidates (e.g., *PICALM*, *BIN1*) whose effects are mediated through tissue-specific expression.[10] Proteome-wide analyses reveal 43 AD-associated proteins, including *TOMM40* and *APOC1*, validated through mediation testing of pQTL effects (63% concordance).[11] These associations were detected due to increased statistical power and reduced multiple testing burden, suggesting that these approaches can aid in identifying the underlying mechanisms of complex traits.[10,11]

Polygenic scores (PGS), which aggregate the effects of multiple common genetic variants identified from GWAS, have historically underperformed in predicting AD risk, as AD is influenced by a combination of genetic, molecular, and environmental factors. The limitations of PGS models for AD prediction are well-documented.[3,12–14] Leonenko *et al* reported that PGS for AD computed using PRS-CS achieved area under the receiver operating characteristic (AUROC) values ranging from 0.55 to 0.75.[12] Their best-performing model, incorporating genome-wide significant single nucleotide polymorphisms (SNPs) and *APOE*, achieved an R^2 of 0.24. Notably, excluding *APOE* from the model reduced the R^2 to 0.06, highlighting the influence of *APOE* on AD risk prediction.[12] Instead, models that incorporate functional information may provide deeper insights into disease etiology and offer more robust tools for individualized risk assessment. Studies developing gene expression risk scores (GeRS) with PGS demonstrated improved predictive performance across phenotypes, underscoring the potential of multimodal approaches in enhancing risk prediction.[15] Furthermore, ensemble machine learning approaches in PGS for SNP selection and modeling have been shown to better account for non-linear and interaction effects, suggesting that incorporating nonlinear modeling techniques with multi-modal data could further enhance prediction accuracy.[16]

This study explores the use of univariate and multivariate approaches, paired with multi-modal data, to improve risk prediction for LOAD in one of the most comprehensive AD datasets currently available. We use univariate approaches—GWAS, TWAS, and PWAS—to identify associations within each molecular layer. These approaches

RESEARCH IN CONTEXT

1. **Systematic review**: We reviewed studies predicting Alzheimer's disease (AD) risk using genome-wide association studies (GWAS), polygenic scores (PGS), and more recent transcriptome-wide (TWAS) and proteome-wide (PWAS) approaches. While PGS offer moderate predictive power, they fail to capture the broader biological complexity of late-onset AD.
2. **Interpretation**: Gene-set enrichment of TWAS and PWAS results revealed enriched biological pathways (cholesterol metabolism, immune signaling, and DNA repair) not captured by GWAS results alone, offering additional mechanistic insights into AD risk. Our integrative risk models (IRMs) combining genetically regulated gene and protein expression using nonlinear machine learning approaches significantly outperform baseline covariates and PGS-based models.
3. **Future directions**: Broader use of multi-omics and clinical data is needed to enhance predictive accuracy and generalizability. Future models should incorporate longitudinal data, nonlinear approaches, and assess performance in multiple populations. Multi-modal modeling offers a promising path toward precision medicine approaches for early detection and intervention in AD.

are valuable for detecting individual signals but do not capture interactions across modalities. To capture complementary biological information beyond genetic data alone, we use multivariate modeling to develop integrative risk models (IRMs) that integrate transcriptomic and proteomic features (Figure 1), enhancing predictive accuracy and enabling a more comprehensive understanding of disease risk. By addressing the limitations of current approaches and exploring the added value of multi-modal molecular data, we aim to advance understanding of AD risk and improve predictive accuracy for clinical and research applications.

2 | METHODS

2.1 | ADSP quality control

The ADSP R4 dataset from the NIAGADS includes WGS data from a globally diverse population of 11,384 individuals diagnosed with AD, 16,685 individuals without AD, and 8292 individuals with uncertain or alternative diagnoses, spanning 40 global cohorts.[17–19] The full ADSP dataset includes data from nearly all previous large-scale AD studies, including the ADNI, ROS/MAP, and the Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4 Study), among several other cohorts, thus creating one of the most comprehensive existing datasets for AD and related dementias.[20–23] Resequencing, variant calling, and various QC measures were conducted in accordance with the standards established by ADSP, ensuring high-quality genomic data.[24] Additionally, case and control phenotypes were harmonized across cohorts by ADSP experts to reduce variability in diagnostic criteria and improve comparability prior to data release.[17,18,24]

For the present study, the initial dataset of 36,361 individuals was filtered based on defined criteria to enable focused analysis of LOAD. We selected cohorts with a mean age of cases greater than or equal to 70 years, thereby excluding early-onset cohorts. Additional considerations included ensuring balanced sex distribution among cases and controls, by ensuring at least 30% female cases, removing cohorts with very low case–control counts (percentage of cases < 15%), and excluding samples with mixed phenotypes such as other dementias. The resulting subset comprised 15,480 participants (6885 cases and 8595 controls) where harmonized binary case and control status for each individual was based on clinical diagnosis which was determined and released by ADSP. [17,18] Principal component analysis (PCA) was performed to assess the genetic similarity of this global population to labeled reference populations by projecting the ADSP participants against the 1000 Genomes reference populations.[7] The full participant counts by cohort after filtering and QC are available in Table S1.

2.2 | Genome-wide association studies (GWAS)

We conducted a GWAS using PLINK v2.0 after performing filtering and performing QC on the ADSP dataset to perform an initial analysis to discern genetic loci associated with AD risk.[25] As part of QC measures prior to performing GWAS, variant names were standardized by filling in missing identifiers. Variants that failed laboratory-based QC filters were removed, along with intentionally duplicated samples. Additional filtering steps included removing variants with a minor allele count (MAC) less than 20, applying a 99% variant call rate threshold, and a 95% sample call rate threshold. Analyses were conducted both with and without applying a minor allele frequency (MAF) threshold of 0.01 to assess sensitivity to rare variant inclusion.[26] PLINK v2.0 supports an additive model for GWAS; the additive model was used for this GWAS.[25] To ensure robustness and account for potential confounders, our models were adjusted for various parameters, including age at diagnosis for cases or age at date of data release for controls, sex, and the first five principal components (PCs) to account for population stratification.[26] Significant loci were identified at a genome-wide significant p-value threshold of $p < 5E\text{-}08$.

2.3 | Transcriptome-wide association studies (TWAS)

To study tissue-specific gene-expression changes associated with AD, we conducted TWAS using PrediXcan and multivariate adaptive shrinkage (MASHR) eQTL models from the Genotype-Tissue

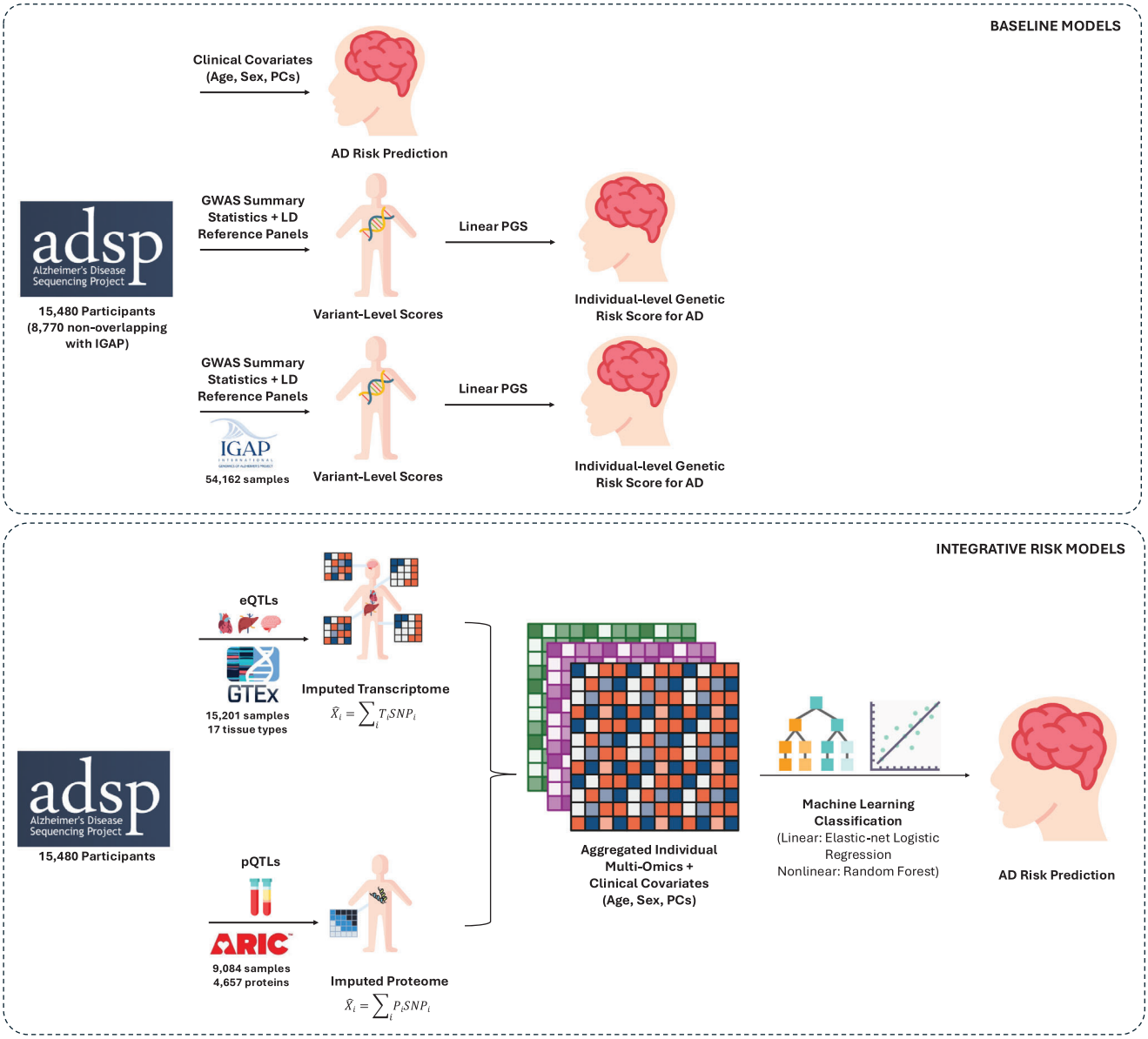

FIGURE 1 Overview of IRM framework. A total of 15,480 participants from ADSP, eQTLs from GTEx, and pQTLs from ARIC were used to impute gene and protein expression. This information is combined with clinical covariates of age, sex, and PCs and then used as features for machine learning classification algorithms. Baseline models of clinical covariates and two separate PGS were generated to evaluate performance against the IRM. ADSP, Alzheimer's Disease Sequencing Project R4; ARIC, Atherosclerosis Risk in Communities study; eQTL, expression quantitative trait loci; GTEx, Genotype-Tissue Expression Project; IRM, integrative risk model; PC, principal component; PGS, polygenic score; pQTL, protein quantitative trait locus.

Expression (GTEx) Project v8, available in PredictDB.[27,28] GTEx eQTLs were derived from a sample group consisting of 84.6% European, 12.9% African American, 1.3% Asian, and 1.1% unknown ancestries, a multi-ancestry cohort similar to that of ADSP.[29] Using this reference, we imputed genetically regulated gene expression (GReX) for all protein coding genes in 17 tissues known to be relevant in AD and neurodegenerative disorders (brain spinal cord cervical c-1, brain cortex, brain substantia nigra, colon sigmoid, liver, vagina, whole blood, brain anterior cingulate cortex BA24, brain frontal cortex BA9, brain hypothalamus, brain caudate basal ganglia, brain nucleus accumbens basal ganglia, brain cerebellar hemisphere, brain putamen basal ganglia, brain cerebellum, brain amygdala, and brain hippocampus).[28–32] Associations with AD status were calculated independently for each of these 17 tissues. Significant genes were determined using a Bonferroni threshold ($p < 0.05/\#$ genes tested) per tissue.

2.4 | Proteome-wide association studies (PWAS)

We then performed PWAS using PrediXcan with the WGS data for the individuals in the ADSP subset. PWAS identifies genetic

associations that may influence complex traits, such as AD, by regulating protein abundance in tissue.[33] Blood plasma-derived protein quantitative trait loci (cis-pQTLs) from the Atherosclerosis Risk in Communities (ARIC) study were used to construct the models.[34] This large bi-ethnic study was made up of 9084 participants, consisting of 7213 European Americans (EA) and 1871 African Americans (AA). PrediXcan PWAS EA and AA models were identified in PredictDB, constructed using ARIC consortium data by utilizing PEERS covariates, expression information from pQTL associations, as well as gene and SNP annotations.[34,35] The genetically regulated components of protein expression were imputed using the EA and AA cohort information, after which PWAS was performed. The resulting PWAS associations were assessed for statistical significance using a Bonferroni threshold ($p < 0.05$/# proteins tested) for each model.

2.5 | Gene-set enrichment analysis

Gene-set enrichment was performed using Enrichr for the significant results from GWAS, TWAS, and PWAS, respectively.[36–38] Enrichment analysis explored the specific pathways and processes associated with the statistically significant genes and proteins from each association study. Pathway results were annotated with the Kyoto Encyclopedia of Genes and Genomes (KEGG) 2021, Reactome 2022, and Gene Ontology (GO) Biological Process 2023 pathways.[37] The significant pathways were identified as having a Fisher's exact test adjusted p-value < 0.05, indicating overrepresentation of these pathways relative to the full set of annotated pathways.[37]

2.6 | Baseline models

2.6.1 | Baseline Polygenic Scores (PGS)

A baseline PGS was created and evaluated on the ADSP dataset, calculated using PRS-CS, a Bayesian regression framework that infers posterior SNP effect sizes under continuous shrinkage (CS) priors using GWAS summary statistics and an external linkage disequilibrium (LD) reference panel.[39] We applied PRS-CS to the ADSP data using GWAS summary statistics generated from 10 iterations of split-sample analyses within ADSP, treating each split as the discovery dataset. To ensure robust and unbiased estimation of the PGS, we performed 10 independent iterations of training and testing. In each iteration, the ADSP dataset was randomly split into 80% training and 20% testing subsets. PRS-CS was applied to the training subset using LD information from a multi-ancestry LD reference panel from 1000 Genomes Project individuals, producing posterior SNP effect sizes for PGS computation.[40] Individual-level PGS values were then calculated using the –score function in PLINK 2.0 by summing the dosages of risk alleles weighted by the PRS-CS posterior effect sizes.[25] After obtaining PGS values for each individual, logistic regression models were trained using the PGS as a single predictor for AD status in the 80% training set of each iteration. Model performance was evaluated on the hold-out 20% set from each iteration, and the final model performance was estimated by averaging metrics such as AUROC, area under the precision-recall curve (AUPRC), sensitivity, and specificity across the 10 test sets. To further assess generalizability, we conducted a nested 80/20 evaluation within the original 20% held-out data from each iteration. This involved splitting the 20% test set into an 80/20 internal validation/test subset, fitting a logistic regression model using the previously computed PGS on the 80%, and evaluating its performance on the internal 20% to simulate model behavior in unseen data. This iterative evaluation approach ensured that the baseline PGS model performance was assessed with reduced risk of overfitting and better reflects the expected predictive power in the ADSP cohort.

A secondary PGS was created using GWAS summary statistics from the International Genomics of Alzheimer's Disease (IGAP) study Stage 1 exclusively with PRS-CS ($n = 54{,}162$).[39,41] A linkage disequilibrium (LD) reference panel of European ancestry individuals from the 1000 Genomes phase 3 project was utilized in the computation.[40] The AD summary statistics from IGAP are the largest set of summary statistics currently available for this phenotype, providing a robust dataset to use for a baseline PGS. The *APOE* ε4 variant (rs429358) was initially excluded from the computation as it was not on the LD reference panel; this variant was added back as a comparison model, using the IGAP GWAS beta as this variant's PGS. To avoid overfitting and biased sampling, the PGS was applied using the PLINKv2.0—score function to individuals in ADSP who do not overlap with IGAP and met phenotyping criteria described above ($n = 8{,}770$). Additional genetic risk scores (GRS) were created using the PLINKv2.0—score function on 1) exclusively SNPs in *APOE* as well as 2) a set of statistically significant AD risk loci from ADVP. The weights were generated from IGAP GWAS beta values and then applied to the ADSP data. All resulting GRS and PGS were fitted with a LASSO regression, using AD case/control status as the outcome and age, sex, and PC1-5 as covariates. The sample was split into a 70% and 30% train/test split with 100 random iterations, balanced by case/control ratio. Performance was assessed with average AUROC, area under the precision-recall curve (AUPRC), F1 score, R^2 calculated using the Matthews Correlation Coefficient (MCC), and balanced accuracy across the iterations.

2.6.2 | Genetic similarity ablation study

We additionally conducted an ablation study to evaluate the impact of population structure on model performance. Two baseline elastic net logistic regression models were constructed to classify risk of AD. The first model included the first 5 PCs derived from genome-wide genetic data, along with age and sex as covariates, to account for genetic similarity and demographic confounding. The second model excluded the PCs, retaining only age and sex as covariates. This comparison allowed us to assess the contribution of genetic population structure to the predictive accuracy and performance of the classification model.

2.7 | Integrative Risk Models (IRMs)

2.7.1 | Elastic net logistic regression models with imputed data modalities

After QC, the ADSP dataset consisted of 15,480 participants from the subset, including 6885 cases and 8595 controls. Participants were selected based on predefined phenotypic criteria, and demographic variables such as age and sex. To account for population structure and minimize confounding effects from population structure, the first five principal components (PCs) were also incorporated as features in the analysis.

Between 1232–10,985 imputed gene and protein expression features were used as inputs to each model, with the exact number depending on the tissue type, as determined by the GTEx project, and protein expression model used from the ARIC study. The full set of models with feature counts are available in Table S2. An elastic net regression model was developed to predict case–control status for each combination of imputed gene expression tissue and imputed protein expression, leveraging a combination of L1 (LASSO) and L2 (ridge) regularization. This approach was utilized due to its ability to handle multicollinearity among predictors while simultaneously performing feature selection and coefficient shrinkage.[42] The dataset was divided into training, validation, and test sets using a 70:15:15 split and scaled column-wise using z-score normalization, by standardizing each feature by removing the mean and scaling to unit variance.[43] The training set was used to fit the model, while the validation set was reserved for hyperparameter tuning. An independent test set was used exclusively for final model evaluation to ensure unbiased performance assessment. Hyperparameter tuning was conducted using grid search with five-fold cross-validation to optimize the L1 ratio, which controls the balance between lasso and ridge penalties, and the regularization strength (C) parameter.[42] The full set of hyperparameters tuned are shown in Table S3. To ensure model convergence during training, the maximum number of iterations was set to 500. To address the high dimensionality of the dataset, two feature selection processes were tested. Least absolute shrinkage and selection operator (LASSO) regression was employed, as well as PCA, which also reduced the dimensionality of the selected features while retaining the most informative components for downstream modeling.[44] PCA was fit on the scaled training set to extract the minimum number of principal components required to explain 80% of the total variance of the dataset. The same transformation was applied to the validation and test sets to ensure consistency across datasets. Of these tests, PCA was determined to most improve performance by reducing more features while maintaining orthogonality in the data.

The comprehensive train-validation-test split framework allowed for robust hyperparameter optimization while mitigating overfitting risks. The final elastic net models were evaluated on the test set to determine predictive performance. The AUROC, AUPRC, F1-score, and balanced accuracy metrics were used to assess the model's ability to discriminate between cases and controls. Ten iterations of each model were run, to get measures of the average and standard deviation of each metric. This approach ensured that the model's performance was robust and generalizable across the diverse tissue-specific datasets analyzed in the study.

2.7.2 | Random forest models with imputed data modalities

To address the limitations of linear modeling approaches in capturing complex, nonlinear relationships between genetic, transcriptomic, and proteomic features, we implemented a series of random forest models.[43] Random forests are well-suited for exploring nonlinear patterns and interactions within high-dimensional datasets.[45] This approach was applied to the same subset of data from ADSP, with the same covariates of age, sex, and first five PCs. The dataset was split into 70:15:15 training, validation, and test sets. The training set was used to fit the random forest model, while the validation set was reserved for hyperparameter tuning. After identifying the optimal hyperparameter combination, the model was evaluated on an independent test set to ensure unbiased performance assessment.

To optimize the random forest models, we employed random search, a strategy that efficiently samples from predefined hyperparameter distributions to identify optimal configurations. This method was chosen for its computational efficiency compared to exhaustive grid search. Hyperparameters of the number of estimators, maximum tree depth, minimum samples required to split a node, minimum samples required at a leaf node, and maximum features considered for a split were all tuned.[46] The full set of hyperparameters tuned are shown in Table S3. Random search was performed over 10 iterations, with five-fold cross-validation used to evaluate model performance for each hyperparameter combination. This iterative process ensured that the selected model configuration for every iteration of each random forest model was optimal for performance and generalizable.

Model performance was measured using AUROC, AUPRC, F1-score, and balanced accuracy. Ten iterations of each model were run, to get measures of the average and standard deviation of each metric. These metrics provided a robust evaluation of the model's ability to discriminate between cases and controls. An illustrative overview of the baseline models and IRM framework is depicted in Figure 1.

3 | RESULTS

3.1 | ADSP study population

Following resequencing, variant calling, and quality control processes adhering to ADSP protocols, phenotype harmonization across cohorts ensured consistent diagnostic criteria for subsequent analyses. PCA revealed that three to four principal components (PCs) accounted for 80% of the explained variance in this subset (Figure S1A). Furthermore, projecting the PCs onto the 1000 Genomes reference population (Figure S1B) showed a similar genetic similarity distribution to the full

TABLE 1 Participant counts from the ADSP subset, stratified by case–control status, biological sex, and genetically inferred ancestry.

Genetic similarity group	Sex	Case–control status	No. of participants
AFR	Female	Control	2157
AFR	Female	Case	1172
AFR	Male	Control	844
AFR	Male	Case	450
AMR	Female	Control	2188
AMR	Female	Case	1270
AMR	Male	Control	957
AMR	Male	Case	634
EUR	Female	Control	1322
EUR	Female	Case	1737
EUR	Male	Control	850
EUR	Male	Case	1332
Other	Female	Control	154
Other	Female	Case	178
Other	Male	Control	110
Other	Male	Case	104
SAS	Female	Control	7
SAS	Female	Case	5
SAS	Male	Control	6
SAS	Male	Case	3

Abbreviations: ADSP, ADSP, Alzheimer's Disease Sequencing Project R4; AFR, African; AMR, Admixed American; EUR, European; SAS, South Asian.

ADSP R4 dataset, suggesting that population-specific information was not lost through the filtering approach.

After filtering the dataset based on the exclusion criteria described in Methods and projecting the PCA of the participants on the 1000 Genomes reference populations, the remaining dataset of 15,480 individuals indicates that sex and case–control status were distributed across five major genetic-similarity groups: African (AFR), Admixed American (AMR), European (EUR), South Asian (SAS), and Other (Table 1). The majority of participants self-identified or genetically clustered as EUR (5241 individuals), followed by AMR (5049), AFR (4623), Other (546), and SAS (21). Within each population group, both cases and controls were represented across sexes, with minor imbalances observed among the control populations. For example, the AFR subgroup had more female cases (1172) than male cases (844), and a higher count of male controls (2157) than female controls (450). Representation from SAS and "Other" groups was limited, comprising fewer than 600 individuals in total, though both case and control statuses were included. Overall, the subset dataset retained diversity across population groups, and ensured sex was balanced within most groups, supporting downstream analyses. The general diversity of ADSP population groups paired with the limited size of some cohorts after filtering justified the cosmopolitan multi-ancestry approach used for the association studies and modeling. The full participant counts by cohort are available in Table S1.

3.2 | Association study and pathway enrichment analyses

We identified 104 genomic, 319 transcriptomic, and 76 proteomic associations with AD via GWAS (p-value < 5E-08), TWAS (p-value < 5.62E-06), and PWAS (p < 5.29E-05), respectively, under genome-wide significance thresholds for GWAS and Bonferroni-significance thresholds for TWAS and PWAS (Figure 2A). Manhattan plots and QQ plots of the GWAS results are available in Figure S2A–C. The significant GWAS risk loci mapped to 12 unique genes by position, 94 unique genes were identified via TWAS, and 17 unique genes identified via PWAS. Of these associations, nine genes and proteins, including *APOE*, *APOC1*, *TOMM40*, *NECTIN2*, *WWC1*, *NRCAM*, *KCNV2*, *CR1*, and amyloid precursor protein (APP) were validated in the Alzheimer's Disease Variant Portal (ADVP).[47] The remaining associations are putatively novel, though some, such as *BLOC1S5-TXNDC5*, have been linked to AD progression in the literature, but have not been previously reported in the ADVP.[47,48] *GLCE* was implicated across all three association studies; this gene is linked to heparan sulfate production and was found in previous studies to be upregulated in AD patients.[49] The full tables of statistically significant associations for each association study are available in Tables S4–S6, and the set of unique annotated genes and proteins and their ADVP status are available in Table S7.

In order to identify the known biologically relevant pathways associated with the statistically significant GWAS, TWAS, and PWAS results, gene-set enrichment analysis was performed using Enrichr for Reactome 2022, KEGG 2021, and gene ontology (GO) 2023 pathways.[37] Significant pathways were identified at an adjusted p-value < 0.05 from the Fisher's exact test; the full pathway results for each association study are available in Table S8. Pathway analysis of GWAS results (Figure 2B) identified significant overrepresentation of lipid metabolism and cholesterol transport pathways, particularly involving *APOC1* and *APOE*. The most enriched pathway was plasma lipoprotein assembly (adjusted p = 2.69E-05), followed by plasma lipoprotein clearance (adjusted p = 1.33E-04) and NR1H3 and NR1H2 regulate gene expression linked to cholesterol transport and efflux (adjusted p = 1.40E-04). GO terms related to phospholipid efflux (adjusted p = 1.23E-05) and high-density lipoprotein particle remodeling (adjusted p = 2.35E-05) were also significantly enriched. These findings underscore the central role of cholesterol homeostasis in Alzheimer's disease pathogenesis, consistent with the known involvement of *APOE* in lipid transport and amyloid-beta metabolism.[7] Additional enriched pathways included cholesterol efflux (adjusted p = 5.66E-05), and broader processes such as cholesterol metabolism (adjusted p = 2.72E-04) and plasma lipoprotein assembly, remodeling, and clearance (adjusted p = 5.49E-04), reinforcing the relevance of lipid regulation mechanisms in genetic susceptibility to AD.[7] Although no pathways reached statistical significance under the adjusted p-value in the TWAS results, several suggestively significant biological

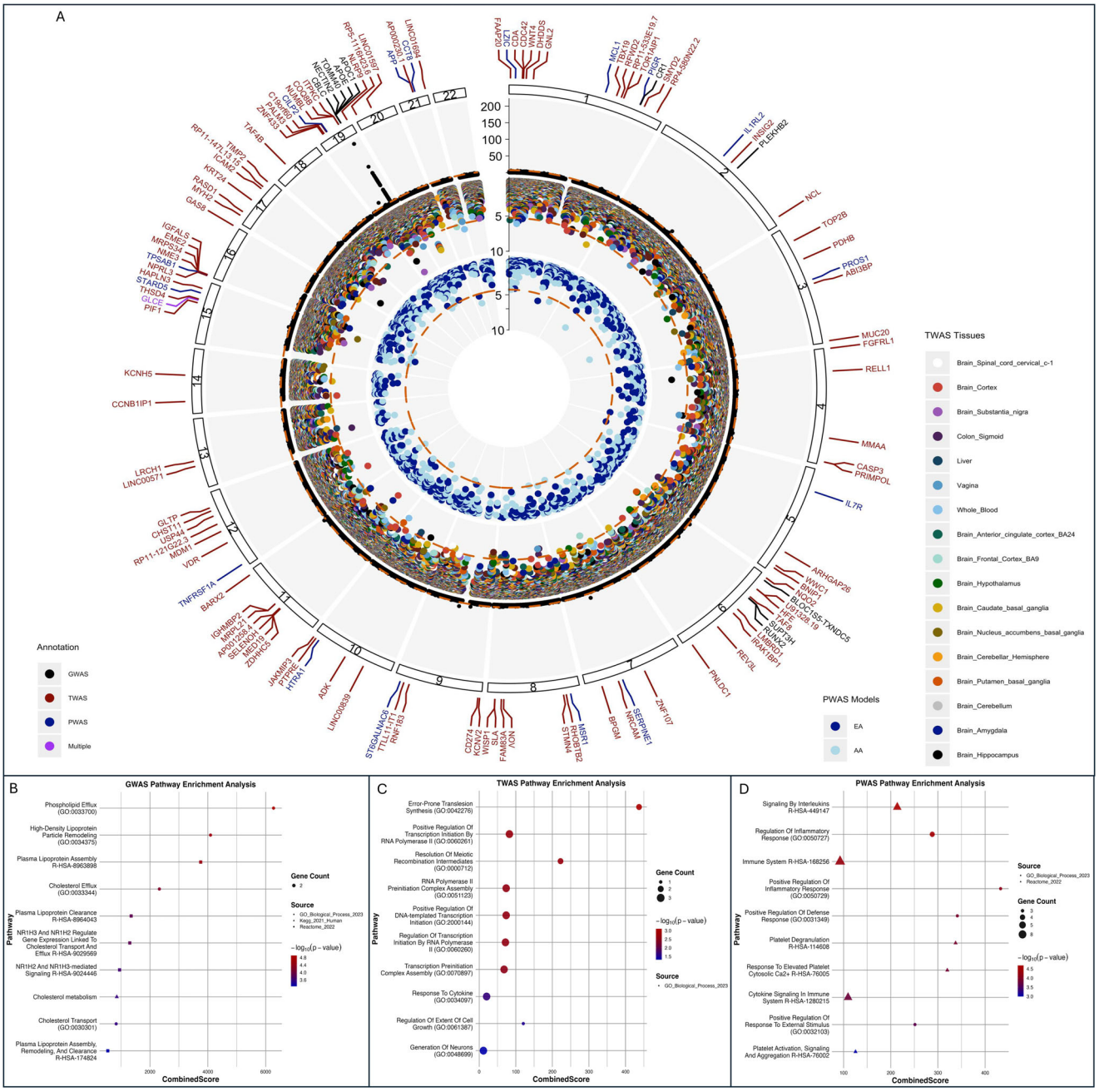

FIGURE 2 Results of association studies and pathway analyses. (A) Circos plot depicting all association study results by tissue and model type. Significant results appear above the dashed orange line denoting Bonferroni significance thresholds. (B–D) KEGG, Reactome, and GO gene-set enrichment results for GWAS, TWAS, and PWAS, colored by *p*-value. Size of the points denotes the number of gene or protein inputs contributing to the pathway. GO, gene ontology; GWAS, genome-wide association studies; KEGG, Kyoto Encyclopedia of Genes and Genomes; PWAS, proteome-wide association studies; TWAS, transcriptome-wide association studies.

processes in neurodegeneration and genomic maintenance were highly ranked (Figure 2C). Among the top-ranked pathways were error-prone translesion synthesis (adjusted $p = 0.2394$) and postreplication repair (adjusted $p = 0.2394$), involving genes such as *REV3L* and *PRIMPOL*, which are involved in DNA damage tolerance[50].

The PWAS pathway results revealed several immune-related pathways with statistically significant enrichment, suggesting a potential link between immune dysregulation and AD (Figure 2D). The top-enriched pathway was Signaling by Interleukins (adjusted $p = 2.07E-05$), driven by proteins including APP, IL1RL2, IL7R, TNFRSF1A, and MCL1, all of which are implicated in inflammatory signaling cascades. The immune system pathway (adjusted $p = 4.92E-05$) was also enriched, encompassing proteins such as APP, IL1RL2, PROS1, PIGR, and CCT8. Several related processes, including cytokine signaling in the immune system, regulation of inflammatory response (adjusted $p = 3.29E-05$), and positive regulation of defense response further

support the hypothesis that systemic immune modulation plays a role in Alzheimer's pathophysiology. Notably, APP, an important contributor to amyloid pathology, appeared consistently across these pathways, underscoring the potential interplay between amyloid biology and immune signaling at the protein level.[11] There were 26 pathways that were found to be significantly or near-significantly enriched across both GWAS and PWAS enrichment results (adjusted $p < 0.1$). Some of these pathways include cholesterol transport (GO:0030301), negative regulation of long-term synaptic potentiation (GO:1900272), and negative regulation of hemostasis (GO:1900047), among others. The full overlapping pathway results are highlighted in Table S8. These overlapping signals lend stronger evidence to the involvement of lipid metabolism, synaptic plasticity, and vascular function in Alzheimer's disease risk. The convergence of enrichment results across genetic and proteomic modalities highlights the biological relevance of these pathways and points toward shared mechanisms contributing to disease susceptibility.

3.3 | IRM model performance

The performance of the elastic-net logistic regression models demonstrated a progressive improvement with the inclusion of additional feature sets, as seen in Table 2. The baseline linear model using only age and sex yielded the lowest performance across all metrics (AUROC = 0.546, AUPRC = 0.459, Balanced Accuracy = 0.496, F1-score = 0.080). The inclusion of PCs as covariates to account for individual-level genetic similarity substantially improved performance (AUROC = 0.636, AUPRC = 0.536, balanced accuracy = 0.604, F1-score = 0.521), indicating that population structure and genetic similarity play a significant role in predictive accuracy. The PGS alone achieved AUROC = 0.608 and R^2 = 0.102 but resulted in a lower AUPRC (0.491) compared to the covariate-only model, suggesting that PGS alone may not capture enough variability in the phenotype. Adding covariates to the PGS generally improved PGS performance (AUROC = 0.623, AUPRC = 0.484, Balanced Accuracy = 0.555, F1-score = 0.681), particularly the R^2, which increased from 0.102 to 0.126. Interestingly, combining PGS with covariates did not yield significant further improvement beyond the covariate-only model, underscoring the potential limitations of PGS in this context. The PGS conducted using scores derived from IGAP demonstrated some performance improvement in AUROC relative to the PGS performed on splits of ADSP, as expected given the large sample size and robust design of the IGAP study. We additionally observed that the GRS consisting of *APOE* alone (AUROC = 0.640, AUPRC = 0.449, balanced accuracy = 0.507, F1-score = 0.058) performed similarly to the PGS of the whole genome including *APOE* (AUROC = 0.641, AUPRC = 0.446, balanced accuracy = 0.511, F1-score = 0.041). This suggests that *APOE* accounts for the majority of the contribution to genetic risk. Performance of the GRS including AD risk loci from ADVP achieved poorer performance (AUROC = 0. 632, AUPRC = 0. 425, balanced accuracy = 0. 505) compared to the model including only *APOE* and the whole genome PRS, suggesting that the ADVP may include some variants that add only noise and may be missing other contributing variants that are present in the whole genome PRS. Overall performance of PGS and baseline models was lacking compared to the multi-omics models.

The integration of imputed multi-omics data showed notable gains, with the transcriptome + covariates model outperforming PGS-based models (AUROC = 0.652, AUPRC = 0.566, balanced accuracy = 0.605). The top five tissues based on median performance were selected, that is, brain cortex, whole blood, brain frontal cortex BA9, brain nucleus accumbens basal ganglia, brain substantia nigra, and performance metrics were averaged across model iterations. The proteome + covariates model performed slightly worse (AUROC = 0.631, AUPRC = 0.548), while integrating both transcriptomic and proteomic data (EA and AA models) produced varying results, with the EA model achieving higher AUROC (0.660) but lower AUPRC (0.521), suggesting differences in predictive power.

The random forest models outperformed the logistic regression models across all feature sets, demonstrating the advantage of non-linear methods in capturing complex relationships within multi-omics data. The transcriptome + covariates model achieved the highest AUROC (0.703) and AUPRC (0.622) among all tested models, showing a substantial improvement over its logistic regression counterpart. The proteome + covariates model also improved in performance (AUROC = 0.680, AUPRC = 0.599), reinforcing the value of protein-level data in predictive modeling. The integration of both transcriptomic and proteomic data (EA and AA models) yielded AUROCs around 0.690 with comparable AUPRC values (~0.613), indicating that multi-omics integration enhances predictive performance regardless of genetic similarity. The average, standard deviation, and confidence interval of all performance metrics across the 10 iterations of each model is shown in Table 2. Boxplots of performance for each tissue-specific model for elastic-net logistic regression are available in Figure S3A, and for random forest in Figure S3B. Across both modeling approaches, the F1-score and balanced accuracy metrics followed similar trends, with random forest models consistently achieving superior results, suggesting improved classification performance. These findings highlight the importance of multi-omics integration and non-linear modeling approaches for improving predictive accuracy in complex phenotypes.

3.4 | Feature importance

To assess feature importance in the best performing RF model, we used the Gini index to quantify the relative contribution of each feature to classification performance. Feature importance across 10 iterations per tissue type was determined. Across the top five tissues, that is, brain cortex, whole blood, brain frontal cortex BA9, brain nucleus accumbens basal ganglia, and brain substantia nigra, the importance of each feature was averaged, as seen in Figure 3A. Our analysis revealed that principal components (PCs) and age were among the most influential predictors. PC4 exhibited the highest importance (0.1387), followed by age (0.0761) and PC1 (0.0750). PC2 (0.0630) and PC5 (0.0414) also demonstrated moderate importance, indicating that

TABLE 2 Performance metrics of the baseline models and IRMs: AUROC, AUPRC, balanced accuracy, and F1 score.

Model	Features	AUROC	AUPRC	Balanced accuracy	F1-Score
Elastic-net logistic regression	Age + sex	0.546 (0.006) (0.542, 0.550)	0.459 (0.007) (0.455, 0.463)	0.496 (0.004) (0.493, 0.498)	0.080 (0.028) (0.062, 0.097)
	Age + sex + PCs (covariates)	0.636 (0.007) (0.632, 0.640)	0.536 (0.008) (0.531, 0.541)	0.604 (0.007) (0.600, 0.608)	0.521 (0.01) (0.515, 0.527)
	ADSP PGS (genome)	0.608 (0.016) (0.604, 0.611)	0.491 (0.015) (0.488, 0.494)	0.540 (0.013) (0.537, 0.542)	**0.691 (0.011) (0.689, 0.693)**
	ADSP PGS + covariates	0.623 (0.017) (0.620, 0.627)	0.484 (0.015) (0.481, 0.487)	0.555 (0.015) (0.552, 0.558)	0.681 (0.012) (0.679, 0.684)
	IGAP PGS + covariates	0.641 (0.001) (0.640, 0.642)	0.446 (0.001) (0.444, 0.449)	0.507 (0.001) (0.506, 0.508)	0.041 (0.002) (0.037, 0.046)
	IGAP PGS no APOE + covariates	0.609 (0.001) (0.607, 0.611)	0.390 (0.001) (0.388, 0.393)	0.500 (6.04E-05) (0.5000, 0.5002)	0.004 (0.0002) (0.003, 0.004)
	IGAP PGS with APOE + covariates	0.641 (0.001) (0.640, 0.643)	0.446 (0.001) (0.444, 0.449)	0.511 (0.001) (0.510, 0.512)	0.041 (0.002) (0.037, 0.046)
	IGAP GRS only APOE + covariates	0.640 (0.001) (0.639, 0.642)	0.449 (0.003) (0.447, 0.452)	0.507 (0.001) (0.506, 0.508)	0.058 (0.003) (0.053, 0.063)
	IGAP GRS AD loci + covariates	0.632 (0.001) (0.630, 0.634)	0.425 (0.001) (0.423, 0.427)	0.505 (0.0003) (0.505, 0.506)	0.033 (0.002) (0.029, 0.036)
	Transcriptome + covariates	0.652 (0.010) (0.649, 0.655)	0.566 (0.013) (0.563, 0.570)	0.605 (0.009) (0.603, 0.608)	0.522 (0.012) (0.518, 0.525)
	Proteome + covariates	0.631 (0.008) (0.628, 0.635)	0.548 (0.011) (0.543, 0.552)	0.602 (0.007) (0.599, 0.605)	0.521 (0.010) (0.517, 0.526)
	Transcriptome + proteome_EA + covariates	0.660 (0.014) (0.654, 0.663)	0.521 (0.020) (0.516, 0.526)	0.574 (0.009) (0.572, 0.577)	0.396 (0.014) (0.392, 0.400)
	Transcriptome + proteome_AA + covariates	0.640 (0.009) (0.638, 0.642)	0.557 (0.013) (0.553, 0.561)	0.602 (0.008) (0.599, 0.604)	0.525 (0.014) (0.521, 0.528)
Random forest	Transcriptome + covariates	**0.703 (0.010) (0.699, 0.707)**	**0.622 (0.012) (0.617, 0.627)**	**0.645 (0.007) (0.642, 0.648)**	0.616 (0.009) (0.613, 0.620)
	Proteome + covariates	0.680 (0.005) (0.677, 0.682)	0.599 (0.007) (0.595, 0.603)	0.636 (0.003) (0.634, 0.637)	0.624 (0.008) (0.620, 0.627)
	Transcriptome + proteome_EA + covariates	0.689 (0.009) (0.685, 0.693)	0.613 (0.014) (0.607, 0.618)	0.637 (0.005) (0.635, 0.639)	0.620 (0.011) (0.615, 0.624)
	Transcriptome + proteome_AA + covariates	0.690 (0.009) (0.686, 0.694)	0.613 (0.014) (0.607, 0.618)	0.637 (0.005) (0.635, 0.639)	0.617 (0.013) (0.612, 0.623)

Note: Values represent the mean performance across iterations with corresponding standard deviations and 95% confidence intervals, with highest performance emphasized in bold.

Abbreviations: ADSP, Alzheimer's Disease Sequencing Project R4; APOE, apolipoprotein E; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic; GRS, genetic risk score; IGAP, International Genomics of Alzheimer's Disease; PC, principal component; PGS, polygenic score.

genetic similarity and demographic factors played a substantial role in the model's predictions.

While gene expression (GE) features showed lower individual Gini importance scores compared to PCs, they remain biologically relevant. *APOC1* (ENSG00000130208.9) (0.0370) had an importance score comparable to PC5, suggesting that tissue-specific molecular variation contributes meaningfully to the model. Similarly, *APOE* (ENSG00000130203.9) (0.0040) and *NECTIN2* (ENSG00000130202.9) (0.0012) displayed non-negligible importance, though their relative contributions were lower. The comparatively modest importance of GE features may be partially attributed to their tissue specificity, as opposed to PCs, which were computed across individuals and therefore capture broader population-wide variance.

When tissue-specific contributions were individually assessed for feature importance (Figure 3B–F), additional genes appeared significant, indicating the importance of tissue-specific variation. Across all five top tissues, PCs and age emerged as the dominant predictors, consistently ranking among the most important features. PC4 was the top predictor in every tissue, with the highest importance observed in the brain frontal cortex (0.1395) and slightly lower but comparable values in the brain cortex (0.1334) and whole blood (0.1351). This component was expected to be the most important as it separated the

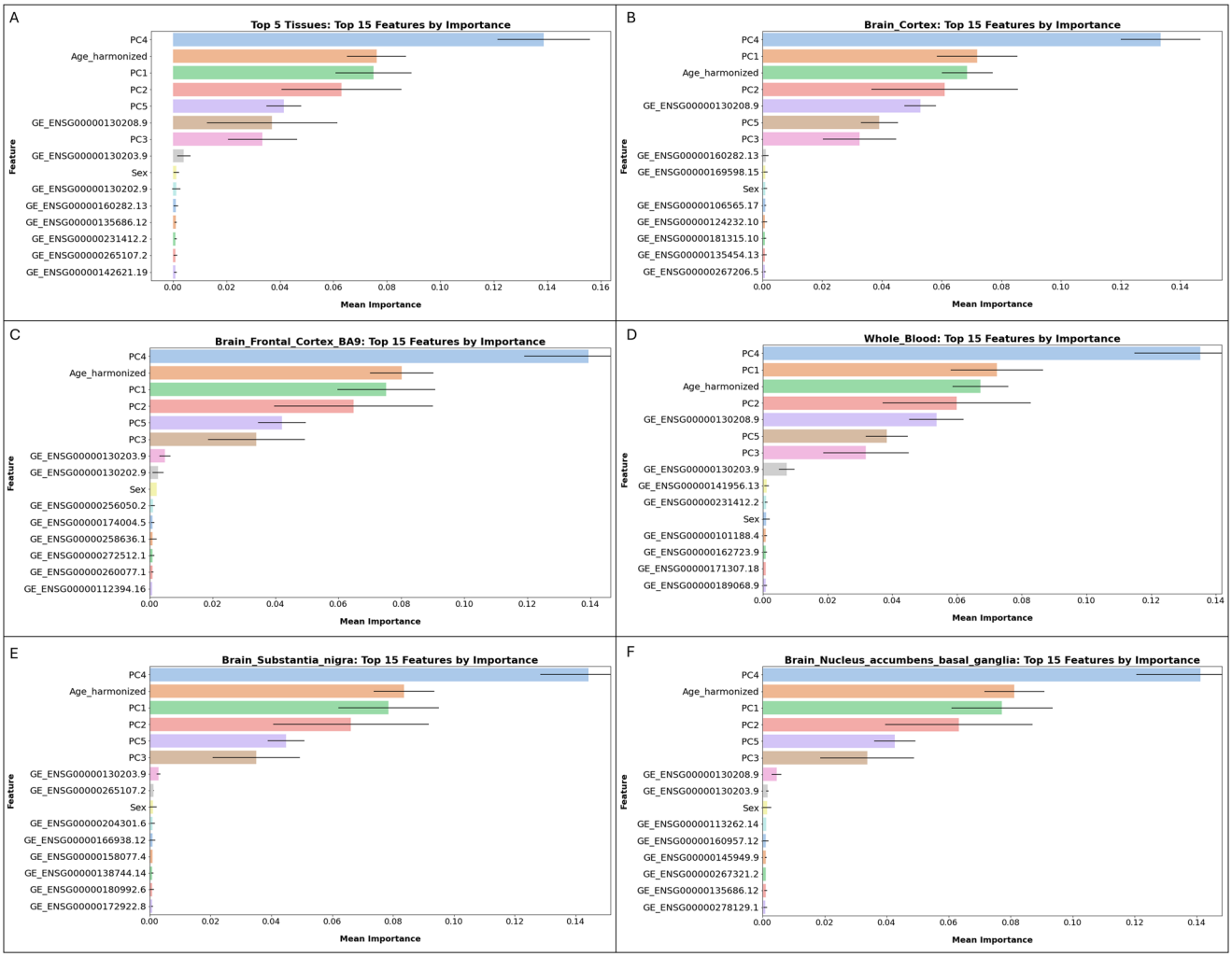

FIGURE 3 Feature importance rankings across top transcriptomic tissues. (A) Mean feature importance values across the top five tissues: brain cortex, whole blood, brain frontal cortex BA9, brain nucleus accumbens basal ganglia, and brain substantia nigra, highlighting consistently predictive genes. (B–F) Tissue-specific feature importance plots showing individual gene contributions to model performance within each tissue.

EUR genetic similarity group making up the greatest proportion of the samples from other populations, as shown in Figure S4. Similarly, PC1, PC2, and PC5 maintained moderate importance across all tissues, suggesting that genetic similarity and demographic factors exert a broad influence on model predictions. The importance of age varied slightly, with the highest impact in the brain frontal cortex (0.0801), followed by the brain cortex (0.0686) and whole blood (0.0673). This suggests that age-related effects may be more pronounced in brain tissues compared to blood, which aligns with known age-related transcriptional changes in the central nervous system.[51] Tissue-specific feature importance for the remaining 12 tissues is shown in Figure S5A–L.

GE feature importance varied significantly between tissues, highlighting the context-dependent role of molecular markers. *APOC1* (ENSG00000130208.9) was among the most important GE features in the brain cortex (0.0529) and whole blood (0.0536) but had lower importance in the brain nucleus accumbens (0.0045) and was absent from the top features in other brain regions. Similarly, *APOE* (ENSG00000130203.9) appeared in multiple tissues but with variable importance, ranging from 0.0073 in whole blood to 0.0016 in the brain nucleus accumbens. Brain substantia nigra prioritized a different set of GE features, with *GJA5* (ENSG00000265107.2) (0.0013) ranking among the top contributors.

4 | DISCUSSION

Our study identified several significant genomic, transcriptomic, and proteomic associations with AD through multi-omics association studies. In addition to well-established loci within genes such as *APOE*, *APOC1*, *TOMM40*, *NECTIN2*, and *CR1*, we also uncovered several putatively novel signals which have been implicated in related biological processes linked to AD progression. Of the 104 genomic, 319 transcriptomic, and 17 proteomic associations, 122 unique genes and proteins were identified as statistically significant. Of these, nine were independently validated in ADVP, providing additional support for the robustness of our association results. Many of the remaining

associations were identified as contributors to AD pathology in published literature and were statistically significant in the GWAS Catalog. These findings reinforce known genetic contributors to AD and suggest new avenues for further biological investigation. For instance, *BLOC1S5-TXNDC5* encodes proteins involved in protein folding and the endoplasmic reticulum stress response. Notably, increased expression of *TXNDC5* has been linked to oxidative stress and neurodegeneration in AD.[48] *RUNX2*, while primarily known for its role in bone development, has also been suggested to have connections to neurodegenerative disease pathways.[52] *CBLC* functions in the ubiquitination and degradation of proteins; processes that, when disrupted, are known contributors to neurodegeneration.[53]

From the TWAS, *PRDM16* has a known role in cellular differentiation in the developing brain, particularly in the production of mature neurons and their specific positions in the neocortex.[54] While it has not been directly linked to AD, it is associated with *PRDM10*, a gene previously listed in the ADVP.[55] *MRPL44*, which functions in mitochondrial protein synthesis, is also associated with *MRPL39* (an ADVP gene).[56] Mitochondrial dysfunction is a consistent hallmark of neurodegenerative diseases, including AD.[57] *PPFIA4* has been linked to synaptic organization and neuronal signaling, processes known to be disrupted early in AD.[58] *CAPN2* encodes a calcium-dependent protease; dysregulation of calpain activity has been reported in neurodegenerative conditions.[59] *BTBD11* may be involved in transcriptional regulation and broader cellular processes via methylation.[60] *PALM3* plays a role in the Toll signaling pathway and in negative regulation of the cytokine-mediated signaling pathway.[61] *PALM2AKAP2*, a paralog of *PALM3*, is a known ADVP gene, and a recent preprint identified associations between *PALM3* and neurodegenerative diseases.[62] *MTG2* contributes to mitochondrial maintenance, reinforcing the broader role of mitochondrial dysfunction in neurodegenerative disease.[56] *SUCO* is involved in maintaining cellular structure, which may impact neuronal integrity.[63] Other significant genes include *RASD1*, involved in neuronal growth and synaptic plasticity, and *ZDHHC5*, a protein palmitoylation gene localized in neuronal structures.[64,65] *ZDHHC5* is linked to neurological disorders and is associated with *ZDHHC20*, an ADVP gene.[65,66]

From the PWAS results, APP (β-amyloid precursor protein) was significantly associated with LOAD, consistent with its central role in β-amyloid plaque formation and disease progression.[7] Several additional proteins identified as significant showed functional relevance despite not being known AD risk genes. For example, PROS1 acts as a ligand for the TAM family of receptor tyrosine kinases, which help mediate anti-inflammatory responses and phagocytic clearance of apoptotic cells and cellular debris, including potentially toxic aggregates like amyloid-β.[67] IL1RAP has been linked to increased amyloid accumulation independently of APOE ε4 status, suggesting alternative mechanisms of amyloid pathology.[68] ANXA7 is a calcium-dependent signaling protein in astrocytes; other annexins have been implicated in neuronal development and AD.[69] HADH, highly expressed in AD brains, binds Aβ and contributes to its cytotoxicity.[70] SERPINE1, encoding PAI-1, is associated with Aβ accumulation and broader neurodegenerative processes.[71] ST6GALNAC6, functionally related to the ADVP gene *ST6GAL2*, has been implicated in glycosylation changes observed in AD.[32,72] *APOE* did not reach Bonferroni-significance in the TWAS or PWAS; this likely reflects limitations of the available tissue-specific references as several tissue models did not include the gene, highlighting a limitation of current single-tissue resources. Concordance across the three association studies was limited; *GLCE* was the sole gene implicated across GWAS, TWAS, and PWAS and is known to be upregulated in AD patients.[49] The lack of overlap likely reflects the differences in reference panels and imputation strategies used for transcriptomic and proteomic inference, underscoring the largely orthogonal signals captured by each approach. The association studies implicate several potential biological mechanisms underlying AD risk that the gene-set enrichment analysis further highlight.

Pathway enrichment analysis of the genomic, transcriptomic, and proteomic association results offers complementary insights into AD biology, with each modality highlighting distinct aspects of disease etiology. The GWAS-based gene set enrichment identified robust associations with lipid metabolism, particularly plasma lipoprotein assembly, remodeling, and cholesterol efflux pathways. These findings are consistent with established roles of apolipoproteins, *APOE* and *APOC1*, in modulating amyloid deposition and neuronal integrity through cholesterol transport and lipid homeostasis.[7] The TWAS-based enrichment analysis did not yield statistically significant pathways after correction, likely due to the relatively large number of tissue-specific genes available for transcriptomic association testing. Nominal enrichments in transcriptional regulation and DNA repair pathways suggest that gene regulation and genome maintenance mechanisms may be enriched in larger AD datasets. Notably, the PWAS analysis revealed a strong signal for immune-related pathways, including interleukin signaling and inflammatory response regulation, implicating immune dysregulation in AD risk.[7,73] This immune enrichment was distinct from the lipid-related processes emphasized in GWAS, suggesting that the proteomic layer may capture systemic or downstream consequences of AD-related perturbations.[73] The 26 pathways enriched across both GWAS and PWAS results include pathways related to cholesterol transport, synaptic potentiation, and vascular dysfunction. Identifying these pathways across multiple modalities provides stronger evidence for their biological relevance to AD and suggests that they represent the most robust signals. These results underscore the value of integrating multi-omics approaches to dissect the complex, multifactorial nature of AD, and point toward the convergence of lipid metabolism, transcriptional regulation, and immune signaling in shaping disease susceptibility.

Building on these molecular insights, our IRMs substantially outperformed traditional approaches. Notably, models incorporating imputed transcriptomic and proteomic expression features consistently outperformed baseline regression models of covariates, as well as PGS across AUROC, AUPRC, and balanced accuracy. The PGS constructed using scores derived from IGAP, one of the largest GWAS efforts in AD, demonstrated modest improvements in AUROC relative to polygenic scores derived from internal ADSP splits. Notably, even this best-case GWAS-driven approach failed to match the predictive performance of the IRMs. These findings underscore the limitations of relying solely

on GWAS-derived polygenic scores, especially across genetically dissimilar populations. The superior performance of the IRMs highlights the added value of incorporating transcriptomic and proteomic information, which provide tissue- and context-specific biological signals absent in PGS. The random forest models further outperformed all linear models, reflecting the utility of non-linear methods in capturing complex, multi-omics relationships. Within the IRMs, PCs emerged as dominant predictors, likely due to capturing large-scale genetic similarity and population structure. GE features also contributed significantly, particularly in tissues such as whole blood and specific brain regions. Tissue-specific variability in GE feature importance suggests that expression regulation in distinct biological contexts meaningfully influences disease risk.

There are several limitations and opportunities for improvement in the development of these IRMs. GREx and GRPx reflect predicted, rather than directly measured molecular phenotypes, based on individual-level genotype data. This was an intentional design choice to enable broader applicability across biobanks where transcriptomic or proteomic profiling may not be available, thereby leveraging scalable prediction models trained on external QTL reference panels to predict heritable risk of developing AD. They may be sensitive to imputation error, especially across different genetic similarity groups, as prediction accuracy is highly dependent on the ancestral composition and size of the reference panels. In the future, this framework can be employed to integrate and assess sequenced multi-omics data rather than relying on imputed information, as datasets become more readily available. Additionally, the tissue-specific nature of GE limits the completeness of prediction models, as not all relevant tissues (e.g., microglia or hippocampal subregions) were captured in the available resources. Moreover, the pQTL reference panel is limited to blood, likely contributing to the observed inflammatory pathway associations in PWAS. While systemic inflammation in blood has been linked to AD and may capture peripheral immune activation, neuroinflammation from glial cells represents a distinct contributor through Aβ plaque formation, that is not captured in blood-based panels.[74] The extent to which peripheral markers reflect brain-specific immune activity remains unclear, underscoring a limitation and a need to integrate brain-derived proteomic and transcriptomic resources to better assess neuroinflammation in AD risk.

Another limitation is the lack of direct biological validation and independent replication of significant TWAS and PWAS associations. While additional computational replication would be valuable before resource-intensive experimental work, suitable datasets presently remain limited, as the ADSP is currently the most comprehensive AD genetic resource and comparable large-scale cohorts with equivalent phenotyping are not yet available. As comparable resources become available, future studies to validate these associations computationally and biologically should be performed. Environmental factors and cellular heterogeneity, which are not reflected in static genomic data, also introduce sources of unexplained variance. Future iterations of these models can account for additional contextual regulation by incorporating additional relevant tissues and integrating longitudinal multi-omics data to better model dynamic biological processes and tissue-specific effects. Moreover, although our models are adjusted for PCs, population stratification remains a potential confounder in genetic risk prediction, and future work should prioritize increasing sample size across populations.

This study demonstrates the clinical and biological value of integrating genetically regulated gene and protein expression with genetic variants for AD risk prediction. Our IRM framework advances beyond traditional PGS approaches by incorporating biologically meaningful, tissue-specific regulatory information. The improved predictive performance, particularly with non-linear modeling, underscores the potential of multi-omics integration to enhance precision medicine strategies. Importantly, the generalizability of our framework offers a scalable pathway for expanding this approach to other complex traits and diseases, ultimately supporting the development of individualized risk prediction tools for clinical decision support.

AUTHOR CONTRIBUTIONS

Rasika Venkatesh: Conceptualization; data curation; analysis; methodology; software; visualization; writing—original draft; writing—review and editing. **Katie M. Cardone**: Data curation; analysis; writing—review; and editing. **Yuki Bradford**: Data curation; analysis; writing—review; and editing. **Anni K. Moore**: Data curation; analysis; writing—review; and editing. **Rachit Kumar**: Data curation; analysis; writing—review; and editing. **Jason H. Moore**: Funding acquisition; supervision; writing—review; and editing. **Li Shen**: Funding acquisition; supervision; writing—review; and editing. **Dokyoon Kim**: Funding acquisition; supervision; writing—review; and editing. **Marylyn D. Ritchie**: Conceptualization; data curation; funding acquisition; resources; supervision; writing—review and editing.

ACKNOWLEDGMENTS

Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging. The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Sequencing for the Follow Up Study (FUS) is supported through U01AG057659 (to Drs. PericakVance, Mayeux, and Vardarajan) and

U01AG062943 (to Drs. Pericak-Vance and Mayeux). Data generation and harmonization in the Follow-up Phase is supported by U54AG052427 (to Drs. Schellenberg and Wang). The FUS Phase analysis of sequence data is supported through U01AG058589 (to Drs. Destefano, Boerwinkle, De Jager, Fornage, Seshadri, and Wijsman), U01AG058654 (to Drs. Haines, Bush, Farrer, Martin, and Pericak-Vance), U01AG058635 (to Dr. Goate), RF1AG058066 (to Drs. Haines, Pericak-Vance, and Scott), RF1AG057519 (to Drs. Farrer and Jun), R01AG048927 (to Dr. Farrer), and RF1AG054074 (to Drs. Pericak-Vance and Beecham). The ADGC cohorts include: Adult Changes in Thought (ACT) (U01 AG006781, U19 AG066567), the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG066506, P30 AG066508, P30 AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20 AG068077, P20 AG068082, P30 AG072958, P30 AG072959), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indiana Memory and Aging Study (IMAS) (R01 AG019771), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD) (U24 AG021886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA- LOAD) (U24 AG056270), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer's Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University Hispanic Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904), the EU Joint Programme—Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research

Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The FUS cohorts include: the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG066506, P30 AG066508, P30 AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20 AG068077, P20 AG068082, P30 AG072958, P30 AG072959), Alzheimer's Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer's Disease Initiative (CuAADI) (3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473, RF1AG015473, R56AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwangju Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hillblom Aging Network (2014-A-004-NET, R01AG032289, R01AG048234), Hussman Institute for Human Genomics Brain Bank (HIHGBB) (R01AG027944, Alzheimer's Association "Identification of Rare Variants in Alzheimer Disease"), Ibadan Study of Aging (IBADAN) (5R01AG009956), Longevity Genes Project (LGP) and LonGenity (R01AG042188, R01AG044829, R01AG046949, R01AG057909, R01AG061155, P30AG038072), Mexican Health and Aging Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, 2R01AG048927), Northern Manhattan Study (NOMAS) (R01NS29993), Peru Alzheimer's Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PR1066) (Wellcome Trust (GR066133/GR080002), European Research Council (340755)), Puerto Rican Alzheimer Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI) (U01AG052410), the Religious Orders Study (ROS) (P30 AG10161, P30 AG72975, R01 AG15819, R01 AG42210), the RUSH Memory and Aging Project (MAP) (R01 AG017917, R01 AG42210Stanford Extreme Phenotypes in AD (R01AG060747), University of Miami Brain Endowment Bank (MBB), University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, R01AG028786), Wisconsin Registry for Alzheimer's Prevention (WRAP) (R01AG027161 and R01AG054047), Mexico-Southern California Autosomal Dominant Alzheimer's Disease Consortium (R01AG069013), Center for Cognitive Neuroscience and Aging (R01AG047649), and the A4 Study (R01AG063689, U19AG010483 and U24AG057437). The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079). Genotyping and sequencing for the ADSP FUS is also conducted at John P. Hussman Institute for Human Genomics (HIHG) Center for Genome Technology (CGT). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U24AG072122) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. Harmonized phenotypes were provided by the ADSP Phenotype Harmonization Consortium (ADSP-PHC), funded by NIA (U24 AG074855, U01 AG068057 and R01 AG059716) and Ultrascale Machine Learning to Empower Discovery in Alzheimer's Disease Biobanks (AI4AD, U01 AG068057). This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations. The ADSP Phenotype Harmonization Consortium (ADSP-PHC) is funded by NIA (U24 AG074855, U01 AG068057 and R01 AG059716). The harmonized cohorts within the ADSP-PHC include: the Anti-Amyloid Treatment in Asymptomatic Alzheimer's study (A4 Study), a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study is funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health-National Institute on Aging, Eli Lilly and Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation and additional private donors, with in-kind support from Avid and Cogstate. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study is funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies are led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School, and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN Studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University

of Southern California. The participants screening for the A4 Study provided permission to share their de-identified data in order to advance the quest to find a successful treatment for Alzheimer's disease. We would like to acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN Studies possible. The complete A4 Study Team list is available on: a4study.org/a4-study-team.; the Adult Changes in Thought study (ACT), U01 AG006781, U19 AG066567; Alzheimer's Disease Neuroimaging Initiative (ADNI): Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California; Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA): 5R37AG015473, RF1AG015473, R56AG051876; the Health & Aging Brain Study—Health Disparities (HABS-HD), supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG054073, R01AG058533, R01AG070862, P41EB015922, and U19AG078109; the Korean Brain Aging Study for the Early Diagnosis and Prediction of Alzheimer's disease (KBASE), which was supported by a grant from Ministry of Science, ICT and Future Planning (Grant No: NRF-2014M3C7A1046042); Memory & Aging Project at Knight Alzheimer's Disease Research Center (MAP at Knight ADRC): The Memory and Aging Project at the Knight-ADRC (Knight-ADRC). This work was supported by the National Institutes of Health (NIH) grants R01AG064614, R01AG044546, RF1AG053303, RF1AG058501, U01AG058922, and R01AG064877 to Carlos Cruchaga. The recruitment and clinical characterization of research participants at Washington University was supported by NIH grants P30AG066444, P01AG03991, and P01AG026276. Data collection and sharing for this project was supported by NIH grants RF1AG054080, P30AG066462, R01AG064614 and U01AG052410. We thank the contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and Psychiatry at Washington University School of Medicine; National Alzheimer's Coordinating Center (NACC): The NACC database is funded by NIA/NIH Grant U24 AG072122. SCAN is a multi-institutional project that was funded as a U24 grant (AG067418) by the National Institute on Aging in May 2020. Data collected by SCAN and shared by NACC are contributed by the NIA-funded ADRCs as follows: P30 AG062429 (PI James Brewer, MD, Ph.D.), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, Ph.D.), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, Ph.D.), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, Ph.D.), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, Ph.D.), P30 AG066519 (PI Frank LaFerla, Ph.D.), P30 AG062677 (PI Ronald Petersen, MD, Ph.D.), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, Ph.D.), P30 AG072946 (PI Linda Van Eldik, Ph.D.), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, Ph.D.), P30 AG066508 (PI Stephen Strittmatter, MD, Ph.D.), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, Ph.D.), P30 AG072931 (PI Henry Paulson, MD, Ph.D.), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, Ph.D.), P20 AG068053 (PI Justin Miller, Ph.D.), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, Ph.D.), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD); National Institute on Aging Alzheimer's Disease Family Based Study (NIA-AD FBS): U24 AG056270; Religious Orders Study (ROS): P30AG10161,R01AG15819, R01AG42210; Memory and Aging Project (MAP—Rush): R01AG017917, R01AG42210; Minority Aging Research Study (MARS): R01AG22018, R01AG42210; the Texas Alzheimer's Research and Care Consortium (TARCC), funded by the Darrell K Royal Texas Alzheimer's Initiative, directed by the Texas Council on Alzheimer's Disease and Related Disorders; Washington Heights/Inwood Columbia Aging Project (WHICAP): RF1 AG054023;and Wisconsin Registry for Alzheimer's Prevention (WRAP): R01AG027161 and R01AG054047. Additional acknowledgments include the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. This work was supported by the National Institutes of Health (HL-169458, U01 AG066833, P30AG073105). R.K. was partially supported by the Training Program in Computational Genomics grant from the National Human Genome Research Institute to the University of Pennsylvania (T32HG000046).

M.D.R., J.H.M., and L.S. were supported by U01 AG066833. L.S. and D.K. were additionally supported by U19 AG074879. All funding bodies had no role in the analysis or interpretation of data in this study.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests and nothing to disclose. Author disclosures are available in the Supporting Information.

CONSENT STATEMENT

All subjects were consented according to ADSP procedures, as described in the ADSP Acknowledgement Statement, prior to data acquisition and release.

ORCID

Rasika Venkatesh https://orcid.org/0009-0003-4471-4291
Marylyn D. Ritchie https://orcid.org/0000-0002-1208-1720

REFERENCES

1. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022;54:412-436. doi:10.1038/s41588-022-01024-z
2. Duchesne S, Rousseau L-S, Belzile-Marsolais F, et al. A scoping review of Alzheimers disease hypotheses: an array of Uni- and multi-factorial theories. *J Alzheimers Dis nd*;99:843-856. doi:10.3233/JAD-230772
3. Karlsson IK, Escott-Price V, Gatz M, et al. Measuring heritable contributions to Alzheimer's disease: polygenic risk score analysis with twins. *Brain Communications*. 2022;4:fcab308. doi:10.1093/braincomms/fcab308
4. Cuyvers E, Sleegers K. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *Lancet Neurol*. 2016;15:857-868. doi:10.1016/S1474-4422(16)00127-7
5. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*. 2006;63:168-174. doi:10.1001/archpsyc.63.2.168
6. Zhu J, Liu X, Yin H, Gao Y, Yu H. Convergent lines of evidence support BIN1 as a risk gene of Alzheimer's disease. *Hum Genomics*. 2021;15:9. doi:10.1186/s40246-021-00307-6
7. Calabrò M, Rinaldi C, Santoro G, Crisafulli C. The biological pathways of Alzheimer disease: a review. *AIMS Neurosci*. 2020;8:86-132. doi:10.3934/Neuroscience.2021005
8. Schlosser P, Zhang J, Liu H, et al. Transcriptome- and proteome-wide association studies nominate determinants of kidney function and damage. *Genome Biology*. 2023;24:150. doi:10.1186/s13059-023-02993-y
9. Jia T, Ma Y, Qin F, Han F, Zhang C. Brain proteome-wide association study linking-genes in multiple sclerosis pathogenesis. *Ann Clin Transl Neurol*. 2023;10:58-69. doi:10.1002/acn3.51699
10. Liu N, Xu J, Liu H, et al. Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease. *PLoS Genet*. 2021;17:e1009363. doi:10.1371/journal.pgen.1009363
11. Hu T, Parrish RL, Dai Q, et al. Omnibus proteome-wide association study identifies 43 risk genes for Alzheimer disease dementia. *Am J Hum Genet*. 2024;111:1848-1863. doi:10.1016/j.ajhg.2024.07.001
12. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun*. 2021;12:4506. doi:10.1038/s41467-021-24082-z
13. Stocker H, Möllers T, Perna L, Brenner H. The genetic risk of Alzheimer's disease beyond APOE ε4: systematic review of Alzheimer's genetic risk scores. *Transl Psychiatry*. 2018;8:1-9. doi:10.1038/s41398-018-0221-8
14. Stocker H, Perna L, Weigl K, et al. Prediction of clinical diagnosis of Alzheimer's disease, vascular, mixed, and all-cause dementia by a polygenic risk score and APOE status in a community-based cohort prospectively followed over 17 years. *Mol Psychiatry*. 2021;26:5812-5822. doi:10.1038/s41380-020-0764-y
15. Pain O, Glanville KP, Hagenaars S, et al. Imputed gene expression risk scores: a functionally informed component of polygenic risk. *Human Molecular Genetics*. 2021;30:727-738. doi:10.1093/hmg/ddab053
16. Elgart M, Lyons G, Romero-Brufau S, et al. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun Biol*. 2022;5:1-12. doi:10.1038/s42003-022-03812-z
17. Wang L-S, Leung YY, Lee W-P, et al. Genomic frontiers in alzheimer's research: a primer on the alzheimer's disease sequencing project (ADSP) and its AI/ML opportunities. *Alzheimers Dement*. 2024;20:e087888. doi:10.1002/alz.087888
18. Kunkle BW, Farrell JJ, Akgun B, et al. Multi-ancestry whole genome sequencing analysis of 36,361 individuals from the Alzheimer's disease sequencing project (ADSP). *Alzheimers Dement*. 2025;20:e092063. doi:10.1002/alz.092063
19. Kuzma A, Valladares O, Greenfest-Allen E, et al. NIAGADS: a comprehensive national data repository for alzheimer's disease and related dementia genetics. Alzheimer's & Dementia 2025;21e70255 . doi:10.1002/alz.70255
20. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64:S161-S189. doi:10.3233/JAD-179939
21. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. *Curr Alzheimer Res*. 2012;9:628-645. doi:10.2174/156720512801322573
22. Mueller SG, Weiner MW, Thal LJ, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*. 2005;15:869-877. doi:10.1016/j.nic.2005.09.008
23. Sperling RA, Rentz DM, Johnson KA, et al. The A4 study: stopping AD before symptoms begin?. *Sci Transl Med*. 2014;6:228fs13. doi:10.1126/scitranslmed.3007941
24. Naj AC, Lin H, Vardarajan BN, et al. Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the Alzheimer's disease sequencing project. *Genomics*. 2019;111:808-818. doi:10.1016/j.ygeno.2018.05.004
25. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4:s13742-015-0047-0048. doi:10.1186/s13742-015-0047-8
26. Turner S, Armstrong LL, Bradford Y, et al. Quality Control Procedures for Genome Wide Association Studies. Current Protocols in Human Genetics 2011;68:1.19.1-1.19.18.doi:10.1002/0471142905.hg0119s68
27. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47:1091-1098. doi:10.1038/ng.3367
28. Barbeira AN, Bonazzola R, Gamazon ER, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol*. 2021;22:49. doi:10.1186/s13059-020-02252-4
29. THE GTEX CONSORTIUM. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318-1330. doi:10.1126/science.aaz1776
30. Bednarska-Makaruk M, Graban A, Wiśniewska A, et al. Association of adiponectin, leptin and resistin with inflammatory markers and obesity in dementia. *Biogerontology*. 2017;18:561-580. doi:10.1007/s10522-017-9701-0

31. Wingo AP, Fan W, Duong DM, et al. Shared proteomic effects of cerebral atherosclerosis and Alzheimer's disease on the human brain. *Nat Neurosci*. 2020;23:696-700. doi:10.1038/s41593-020-0635-5
32. Serrano-Pozo A, Li H, Li Z, et al. Astrocyte transcriptomic changes along the spatiotemporal progression of Alzheimer's disease. *Nat Neurosci*. 2024;27:2384-2400. doi:10.1038/s41593-024-01791-4
33. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biology*. 2020;21:173. doi:10.1186/s13059-020-02089-x
34. Zhang J, Dutta D, Köttgen A, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet*. 2022;54:593-602. doi:10.1038/s41588-022-01051-w
35. Mi S, Protein prediction models—ARIC | PredictDB. PredictDB—Protein Prediction Models—ARIC 2022. https://predictdb.org/post/2022/11/14/protein-prediction-models/ (accessed August 6, 2024)
36. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;14:128. doi:10.1186/1471-2105-14-128
37. Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with enrichr. *Current Protocols*. 2021;1:e90. doi:10.1002/cpz1.90
38. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016;44:W90-W97. doi:10.1093/nar/gkw377
39. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;10:1776. doi:10.1038/s41467-019-09718-5
40. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74. doi:10.1038/nature15393
41. Lambert J-C, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. 2013;45:1452-1458. doi:10.1038/ng.2802
42. Park H, Konishi S. Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*. 2016;86:1450-1461. doi:https://doi.org/10.1080/00949655.2015.1073290
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 2011;12:2825-2830.
44. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. *Nat Rev Methods Primers*. 2022;2:1-21. doi:10.1038/s43586-022-00184-w
45. Breiman L. Random Forests. *Machine Learning*. 2001;45:5-32. doi:10.1023/A:1010933404324
46. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*. 2020;415:295-316. doi:10.1016/j.neucom.2020.07.061
47. Kuksa PP, Liu C-L, Fu W, et al. Alzheimer's Disease variant portal: a catalog of genetic findings for Alzheimer's disease. *J Alzheimers Dis*. 2022;86:461-477. doi:10.3233/JAD-215055
48. Horna-Terrón E, Pradilla-Dieste A, Sánchez-de-Diego C, Osada J. TXNDC5, a newly discovered disulfide isomerase with a key role in. *Cell Physiology and Pathology International Journal of Molecular Sciences*. 2014;15:23501-23518. doi:10.3390/ijms151223501
49. Pérez-López N, Martín C, García B, et al. Alterations in the expression of the genes responsible for the synthesis of heparan sulfate in brains with Alzheimer disease. *J Neuropathol Exp Neurol*. 2021;80:446-456. doi:10.1093/jnen/nlab028
50. Mellor C, Nassar J, Šviković S, Sale JE. PRIMPOL ensures robust handoff between on-the-fly and post-replicative DNA lesion bypass. *Nucleic Acids Research*. 2024;52:243-258. doi:10.1093/nar/gkad1054
51. Saetre P, Jazin E, Emilsson L. Age-related changes in gene expression are accelerated in Alzheimer's disease. *Synapse*. 2011;65:971-974. doi:10.1002/syn.20933
52. Nakatsu D, Kunishige R, Taguchi Y, et al. BMP4-SMAD1/5/9-RUNX2 pathway activation inhibits neurogenesis and oligodendrogenesis in Alzheimer's patients' iPSCs in senescence-related conditions. *Stem Cell Reports*. 2023;18:688-705. doi:10.1016/j.stemcr.2023.01.004
53. Schmidt MF, Gan ZY, Komander D, Dewson G. Ubiquitin signalling in neurodegeneration: mechanisms and therapeutic opportunities. *Cell Death Differ*. 2021;28:570-590. doi:10.1038/s41418-020-00706-7
54. Su L, Lei X, Ma H, Feng C, Jiang J, Jiao J. PRDM16 orchestrates angiogenesis via neural differentiation in the developing brain. *Cell Death Differ*. 2020;27:2313-2329. doi:10.1038/s41418-020-0504-5
55. Nazarian A, Yashin AI, Kulminski AM. Genome-wide analysis of genetic predisposition to. *Alzheimer's Research & Therapy*. 2019;11:5. doi:10.1186/s13195-018-0458-8
56. Del Giudice L, Alifano P, Calcagnile M, et al. Mitochondrial ribosomal protein genes connected with Alzheimer's and tellurite toxicity. *Mitochondrion*. 2022;64:45-58. doi:10.1016/j.mito.2022.02.006
57. Yang H-M. Mitochondrial dysfunction in neurodegenerative diseases. *Cells*. 2025;14:276. doi:10.3390/cells14040276
58. Reid AN, Jayadev S, Prater KE. Microglial responses to alzheimer's disease pathology: insights from "omics" studies. *Glia*. 2025;73:519-538. doi:10.1002/glia.24666
59. Wang L-H, Lin C-Y, Lin Y-M, et al. Calpain-2 mediates MBNL2 degradation and a developmental RNA processing program in neurodegeneration. *J Neurosci*. 2022;42:5102-5114. doi:10.1523/JNEUROSCI.2006-21.2022
60. Smith RG, Hannon E, De Jager PL, et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimers Dement*. 2018;14:1580-1588. doi:10.1016/j.jalz.2018.01.017
61. Chen X, Wu X, Zhao Y, et al. A novel binding protein of single immunoglobulin IL-1 receptor-related molecule: paralemmin-3. *Biochemical and Biophysical Research Communications*. 2011;404:1029-1033. doi:10.1016/j.bbrc.2010.12.104
62. Higginbotham L, Carter EK, Dammer EB, et al. Unbiased classification of the elderly human brain proteome resolves distinct clinical and pathophysiological subtypes of cognitive imparitment. *Neurobiology of Disease*. 2023;186:e106286. doi:10.1016/j.nbs.2023.106286
63. Sha Z, Sha L, Li W, et al. Exome sequencing identifies SUCO mutations in mesial temporal lobe epilepsy. *Neuroscience Letters*. 2015;591:149-154. doi:10.1016/j.neulet.2015.02.009
64. Greenwood MP, Greenwood M, Mecawi AS, Antunes-Rodrigues J, Paton JFR, Murphy D. Rasd1, a small G protein with a big role in the hypothalamic response to neuronal activation. *Mol Brain*. 2016;9:1. doi:10.1186/s13041-015-0182-2
65. Woodley KT, MO Collins. Regulation and function of the palmitoyl-acyltransferase ZDHHC5. *The FEBS Journal*. 2021;288:6623-6634. doi:10.1111/febs.15709
66. Chauhan G, Adams HHH, Satizabal CL, et al. Genetic and lifestyle risk factors for MRI-defined brain infarcts in a population-based setting. *Neurology*. 2019;92:e486-503. doi:10.1212/WNL.0000000000006851
67. Katzeff JS, Bright F, Lo K, et al. Altered serum protein levels in frontotemporal dementia and amyotrophic lateral sclerosis indicate calcium and immunity dysregulation. *Sci Rep*. 2020;10:13741. doi:10.1038/s41598-020-70687-7
68. Ramanan VK, Risacher SL, Nho K, et al. GWAS of longitudinal amyloid accumulation on 18F-florbetapir PET in Alzheimer's disease implicates microglial activation gene IL1RAP. *Brain*. 2015;138:3076-3088. doi:10.1093/brain/awv231
69. White ZB, Nair S, Bredel M. The role of annexins in central nervous system development and disease. *J Mol Med*. 2024;102:751-760. doi:10.1007/s00109-024-02443-7

70. Yan SD, Shi Y, Zhu A, et al. Role of ERAB/l-3-Hydroxyacyl-coenzyme a dehydrogenase type II activity in Aβ-induced cytotoxicity *. *Journal of Biological Chemistry*. 1999;274:2145-2156. doi:10.1074/jbc.274.4.2145
71. Kutz SM, Higgins CE, Higgins PJ. Novel combinatorial therapeutic targeting of PAI-1 (SERPINE1) gene expression in Alzheimer's disease. *Mol Med Ther*. 2012;1:106. doi:10.4172/2324-8769.1000106
72. Yamagata H, Uchida S, Matsuo K, et al. Altered plasma protein glycosylation in a mouse model of depression and in patients with major depression. *Journal of Affective Disorders*. 2018;233:79-85. doi:10.1016/j.jad.2017.08.057
73. Duggan MR, Peng Z, Sipilä PN, et al. Proteomics identifies potential immunological drivers of postinfection brain atrophy and cognitive decline. *Nat Aging*. 2024;4:1263-1278. doi:10.1038/s43587-024-00682-4
74. Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement*. 2018;4:575-590. doi:10.1016/j.trci.2018.06.014

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Venkatesh R, Cardone KM, Bradford Y, et al. Integrative multi-omics approaches identify molecular pathways and improve Alzheimer's disease risk prediction. *Alzheimer's Dement*. 2025;21:e70886. https://doi.org/10.1002/alz.70886